Blackboard Agents For Standard Arabic Language Tokenization And Parsing

Dr.Abdul Kareem Murhij Radhi, College of Information Engineering, Communication and Information Department, Nahrian University, Baghdad,Jadriyia- Iraq.

E-mail: kareem_m_radhi@yahoo.com

Abstract

The Processing of the Arabic language is a difficult mission comparing it with other languages. Because Sentences in Arabic language are complex, longer than the others in various languages, and have difficult structure with lattices. The syntactic structure of sentences parts may be missing, affecting the orders of words and phrases. Parsing Arabic sentences can be done via several techniques which are started with top-down and or bottom-up parsing. Recursive Transition Network one of the most famous techniques for parsing sentences. It is a finite transition automaton of limited states. This paper presents a new method in processing the first stage from Arabic language processing which is a syntax analysis by using Transition Networks techniques.

This technique presents morphological analysis for sentence tokens. Prolog version 2 where used for execution parsing, since it is more consistence for natural languages representation and processing.

Keywords: Pos, Morphemes, Parsing, Tagging, Precision, Recall.

الخلاصة

أن معالجة اللغةِ العربيةِ مهمّةِ صعبةِ مُقَارَنَة باللغاتِ الأخرى،لأن الجُمَل في اللغةِ العربيةِ معقدة وأطول منها في اللغاتِ المُخْتَلِفةِ، ولَها تركيبُ صعبُ متداخل ومتشابك. أن التركيب أو البنية النحوية لأجزاءِ من الجُمَلِ في اللغة العربية قَدْ يَكون مفقودَ أو غير متكامل، والذي يُؤثّرُ ويتأثر بترتيب الكلماتِ والعباراتِ. أن إعراب الجُمَلِ العربيةِ يُمْكِنُ أَنْ يتم عن طريق عِدّة تقنيات التي تَبْدأُ من الأعلى للأسفل أَو من الأسفل للأعلى.

أن الإنتقالُ التكراريُ هو أحد التقنياتِ الأكْثَر شَهْرَةِ لإعراب الجُمَلِ. وهي إنتقال آلي محدود لحالات محددة. يقدم هذا البحث أستخدام طريقة جديدة في معالجة المرحلة الأولى من معالجة اللغة العربية ألا وهي التحليل القواعدي لجمل اللغة العربية المختلفة وذلك بأستخدام تقنية شبكات التحول أو الأنتقال. تستثمر هذه الطريقة تقنية التحليل الصرفي لمفردات الجملة . أستخدمت لغة برولوك نسخة الثانية في تنفيذ عملية الأعراب لكونها اللغة الأكثر ملائمة في تمثيل و معالجة اللغات الطبيعية.

1. INTRODUCTION

Arabic language has 28 letters , 25 of them consonants and three vowels"ي","و","أ" which can be short or long 12 of them are unique to Arabic language, which does not have any corresponding English letters language as "خ","ح","ص","ض" consonants and difficult foreigners to pronounce exactly. In addition the letters are divided into categories. The grammatical system of the Arabic language is based on a root-and-pattern structure and considered as root based language with more than 10000 roots. The Arabic language has received little attention from researches working in the fields of computational linguistics and natural language processing [1].

2- TAGGING ARABIC TOKENS

POS tagging (POS) is considered as one of the basic tools and components necessary for any robust Natural Language Processing infrastructure of a given language [3]. It is needed in many fields of linguistic processing, starting from the simpler ones as text phrasing and alignment, to the more elaborate ones as syntax and semantic analysis and ending up with linguistic processes that is heavy as machine translation. Moreover, POS tagging is also considered as first stage for analyzing and annotating corpora [3].Many taggers have been developed for different languages and used to build various kinds of applications. These applications include speech synthesis, natural language parsing, information retrieval and information extraction [2].

Arabic POS tagging (APOS) is not an easy task due to the high ambiguity results from the absence of diacritics and also from the complexity of the Arabic morphology. Consider the following example: "عالم علم رجلا" (a scholar taught a man). Each word in the above example has more than one morphological analysis. The APOS tagger is responsible for assigning to each word the most appropriate morphological tag.

Part-of-speech tagging consists of assigning to each word of a sentence a tag which indicates the function of that word in that specific context. In Existing Natural Language Processing (NLP) literature, there are many methods that can be classified in three groups: [3]

- Linguistic approach consists of coding the necessary knowledge in a set of rules written by linguist (like the pioneer TAGGIT).
- Statistical approach requires much less human effort, successful model during the last years Hidden Markov Models and related techniques have focused on building probabilistic models of tag transition sequences in sentence.
- The third family use learning algorithms that acquire a language model from a training corpus use an example-based learning technique and a distance measure to decide which of the previously learned examples is more similar to the word to be tagged [6].

3- DEFINITION OF ARABIC LANGUAGE

Arabic language is one of the most popular languages in the world, it is the official language of twenty two Middle East and African countries, and is spoken by more than 200 millions of people all over the world [1] .

Modern Standard Arabic is an adapted from Classical Arabic, which is used in books, newspapers, on television and radio, in the mosques, and in conversation between educated Arabs from different countries [9].

A. Arabic Words Classification

Arabic grammarians traditionally classify words into three main categories: nouns, verbs, and particles. All verbs in Arabic and most of the nouns are derived from the root verbs. These categories are also divided into subcategories, which collectively cover the whole of the Arabic language. These categories are:

A-1. Noun:

A noun in Arabic is a name or a word that describes a person, thing, or idea; the linguistic attributes of nouns are (Gender, Number, Person, Case, and Definiteness).

A-2. Verbs:

Verbs indicate an action, although the more on action and aspects are different. Verb categories are divided into subcategories such as Perfect, Imperfect, and Imperative. The verbal attributes are (Gender, Number, Person...).

B. Arabic Sentence Structure

A Text in Arabic language is composed of set of
Sentences, these Sentences might be a verbal Sentence (جملة فعلية), which has the structure verb-subject-object, and must start or with a verb, it might be a nominal Sentence (جملة أسمية), which must start with a noun [8] .

In each case the sentence is either simple or compound. The difference between the simple sentence and the compound sentence is that the former does not have a complementary that could occur at the end of the sentence [1].

4. Specification of the sentence structure (model architecture):

The sentence in the Arabic language is either nominal like in "الشمس ساطعة" (the sun is bright) or verbal like in "يلعب الأطفال الكرة" (the children play the ball).

Each of them may have different forms and styles. A list of more than 100 ways of common grammatical structures in the Arabic language has been surveyed [12]. It covers the general syntactical analysis and detailed morphological analysis of the nouns and verbs.

4.1. Nominal sentences:

Different forms of formulation have been identified for nominal sentences. They can be represented by the graph (Fig. 1) in terms of sequences, where V, N and P respectively denote Verb, Noun and Particle. S and E are special states, used to represent the start and the end of the nominal phrase. Notice that a loop on a state indicates certain number of repetitions of this symbol and an arrow between two sates, means that first one may be followed by the second one depending on its direction.

4.2 Verbal Sentences:

The verbal sentence is the sentence that begins with the verb followed by the subject and then the predicate, for Example"كتب الطالب الدرس" The verb (كتب) (/kataba/ = he wrote = past) The subject(الطالب = The student) The Accusative Object الدرس) (/addarsa/ = the lesson) When the subject is unknown, we change the verb form, for example:كتب(/kataba/) (Active Verb)كتب (/kutiba/) (Passive verb). We say: (/kutiba-d-darsu/) (كتب الدرس) "the lesson is written" in past tense, in the present tense : (/yuktabu-d-darso/) (يكتب الدرس) figure.2 describe the structure of Verbal Sentences. There are two types of verbs:

4.2.1 The intransitive Verb:

That needs only his subject, for example: (جاء الولد) "the boy comes".

4.2.2 The transitive Verb:

That needs his subject and an accusative object, (المفعول به), for example: (كتب التلميذ الدرس) For the three elements of verb, subject and object, there are different word orders for Subject and Object [9]:

Verb + Subject + Object**:**

Example: (أكل الولد التفاحة: The boy ate the apple**)**. While:Verb + Object + Subject**:** example: **أكل التفاحة الولد) :(**The apple ate the boy)[1].

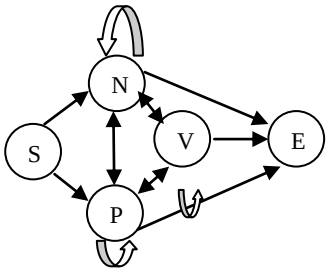


Figure [1] Structure of Arabic Nominal Sentence

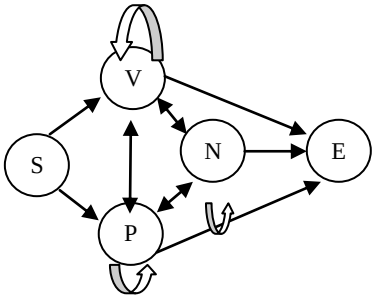


Figure [2] Structure of Arabic Verbal Sentence

5. Morphemes

In lexical module the input Token assign into specific morphemes. If the morpheme can function alone, such as word س (engineer), it is called a free morphed. Other morphemes cannot be used by themselves, such as the general plural ending نو and the letters in the word ن (engineers). Such morphemes are called 'bound'. Bound morphemes, in Arabic, serve as additions at the beginning or ending of a stem. Using the definitions of free and bound morphemes, a word can be defined as a single free morpheme, and can be inflected word can be defined as a complex form which is a single free morpheme combined with one or more bound morpheme[11]. The second lexical analysis task is to assign a suitable symbol to each lexeme. To achieve this task we first have to suggest a symbol (token) to each group of the lexemes, where each group has a common parsing behavior [10]. Moreover Arabic morphemes depends mainly on morphemes weight (الميزان الصرفي") :" فَعَلَ,فَعُول,فاعِل,مِفْعال,مِفْعَل,مَفاعِيلْ,فَعْلَلَ,أَفْتَعَلَ,أَسْتَفْعالْ,مَفْعُوْلْ,فِّعْلِل "

6. Description of the tagging system:

Investigating the principle aspects of Arabic morphology and grammar. The following is a brief review of those aspects. The Arabic verbal structures are composed of three classes: noun (أسم), verb (فعل) and that we will call particle (حرف) [1].

6.1. Noun:

It is either a name or a word that describes a person, thing or idea. It could be definite or indefinite and can be subcategorized by the person (narrator, interlocutor and absent), number (Singular, Dual, Plural), gender (Masculine, Feminine) and grammatical cases (nominative" "الرفع", accusative "النصب ", genitive "الجر "). Figure.3 gives a main classification of the noun and its prominent ramifications.

6.2. Verb:

It is a word that denotes an action and could be combined with some particles. In term of tense (Fig. 2), the verb could be past (perfect ماضي), present (imperfect مضارع) or (imperative أمر) . A future verb tense exists, but it is a derivative of the present tense that you achieve by attaching a prefix to the present tense of the verb.

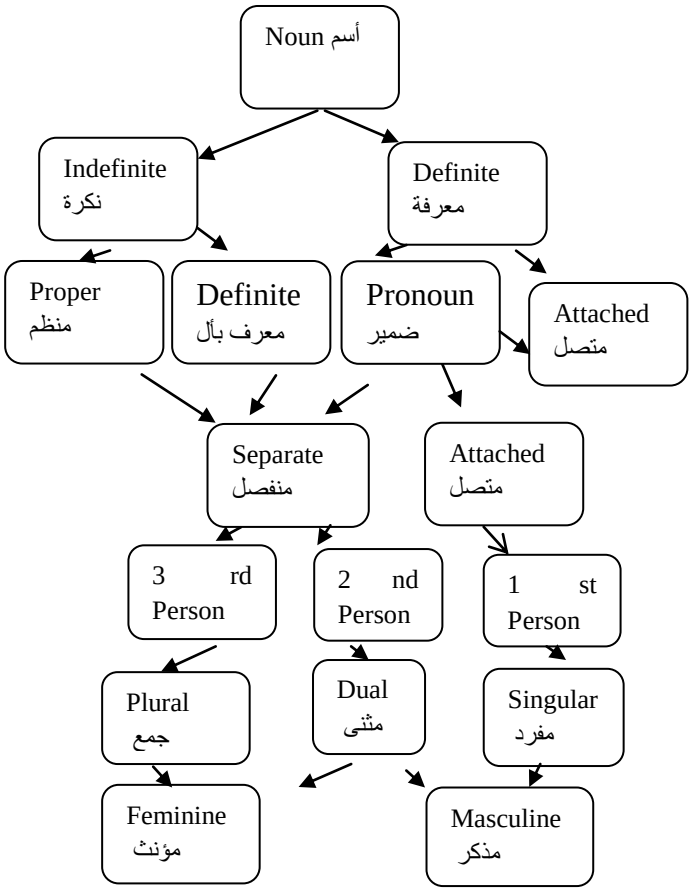


Figure [3] Structure of Noun in Arabic language

Particles can be added as prefixes and/or suffixes indicating the number, gender and person of the subject, like for example: (he says قال), and (she saidقالت), يقولان، يقولون، تقولان، (they say). Three moods are possible for verbs: Indicative " الرفع", subjunctive "'النصب" and jussive"الجزم".

6.3. Particle:

This class includes everything that is neither a verb nor a noun. It contains for example, genitive letters "جر", prepositions of coordination, conjunction as well as the functional words like " أن وأخواتها و كان وأخواتها" which influence the upcoming words analysis. The Particle class includes: Prepositions, Adverbs, Conjunctions, Interrogative Particles, Exceptions and Interjections [7].

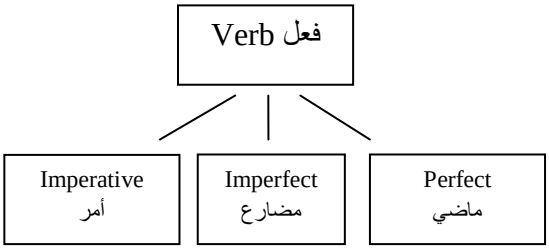


Figure [4] Structure of Arabic Verbs

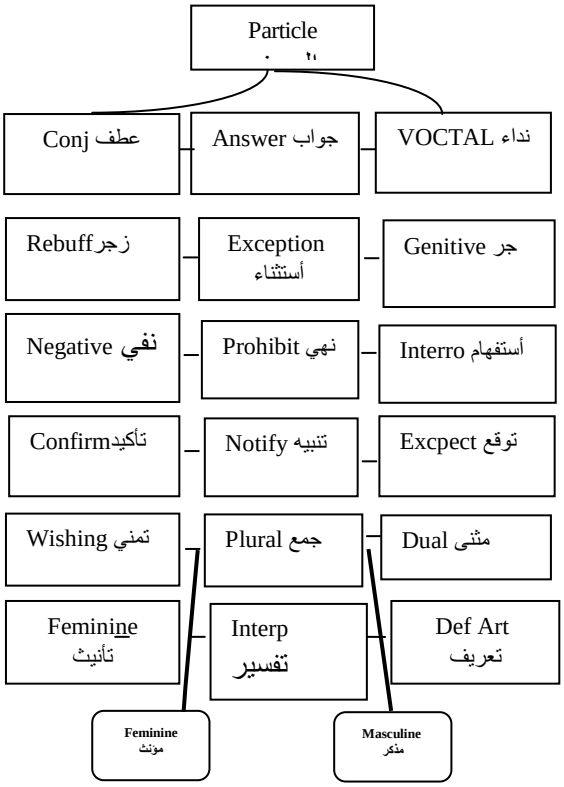


Figure [5] Structure of Arabic Particles

7. PREVIOUS APPROACHES:

7.1. RULES-BASED TAGGING

Consists of developing a knowledge base of rules written by linguists to define precisely how and where to assign the various POS tags.

Several signs in Arabic language that indicate the category of word. One of them is the affix. Some affixes are proper to verbs; some are proper to nouns; and some others are used with verbs and nouns. Another, important sign in Arabic language is the pattern, which is an important guide in recognizing the word category.

Several grammatical rules gives some signs to distinguish between type of word and others signs are deduced from others features (number, gender, preposition and conjunction ...etc. During tagging process, the context and word form features are looked up for each word in the text. Information about surrounding words is should be used [5].

7.2. Statistical approach:

Consists of building a trainable model and to use previously-tagged corpus to estimate its parameters. Once this is done, the model can be used to automatically tagging other texts. Successful statistical taggers were built during the last years and are mainly based on Hidden Markov Models (HMMs) [2].

7.3. Memory-Based Learning

Memory-based learning is a supervised classification-based learning method. A vector of feature values is associated with a class by a classifier that lazily extrapolates from nearest neighbours selected from all stored training examples. Memory-based learning is a direct descent of K-Nearest Neighbour (K-NN) algorithm, it use complex data structure and different speedup optimization from the K-NN. During learning a data base of instances is build with a memory-based learning algorithm IB 1-IG. An instance consists of a fixed-length vector of n feature-value pairs and an information field containing the classification of that particular feature-value vector. The similarity between a new instance x and a memory instance y is computed with a distance metric $\Delta(x,y)$ [1]. The tag of x is then determined by assigning the most frequent category within the k most similar example of x as shown in equation [1].

$$\Delta(X,Y) = \sum \alpha_i \, \delta(x_i,y_i) \text{ ------- (1)}$$

Where α_i is the weight of i-th attribute and

$$\delta(x_i,y_i) = 0 \text{ if } x_i=y_i$$

$$= 1 \text{ if } x_i \neq y_i$$

During tagging process, the context and word form features are looked up for each word in the text. Information about surrounding words is used, two words of the right context and two words of the left context [6].

7.4. Hybrid Method

Consists in combining rule-based approach with a statistical one. Most of the recent study uses this approach as it gives better results.

A memory-based learning system contains two components: i) a learning component which is memory storage is done without abstraction or restructuration. ii) A performance component that does similarity-based classification. The idea, in the proposed method is to apply rules (analyzing the affixes of the word and analyzing its patterns) to determine the tag type of each word in a sentence and to refer to memory-based to check whether it is an exceptional case, or not. Applying rules to predict a tag T_i for a word W_i , the predicted tag Ti is compared with the correct tag in the training phase. In case of no equality, it is considerate as an exception and the type of error is determined according to correct tag and the predicted tag. For each rule the number of exceptional cases is stored in library. Figure [1] shows the structure of the Arabic hybrid tagging model. During classification Firstly, the rules are applied to determine the tag and it is checked as an exceptional case of rules. Secondly, it is presented to memory based reasoning, its similarity to all examples in memory is computed using a similarity metric and the tag is determined again [6].

8. Proposed system

Before starting with training phase, the proposed system first of all preprocessing words via two modules: lexical module and syntax module. The first module processes the words, finds its stems, separates it from prefixes and suffixes, and assigns the filtered words to specific tokens. Syntax module receives tokens, finds the best grammar for the given sentence of the tokens using Recursive Transition Network (RTN). Figure(6), (7) and figure(8) summarize first module, while figure (9) presents the second module.

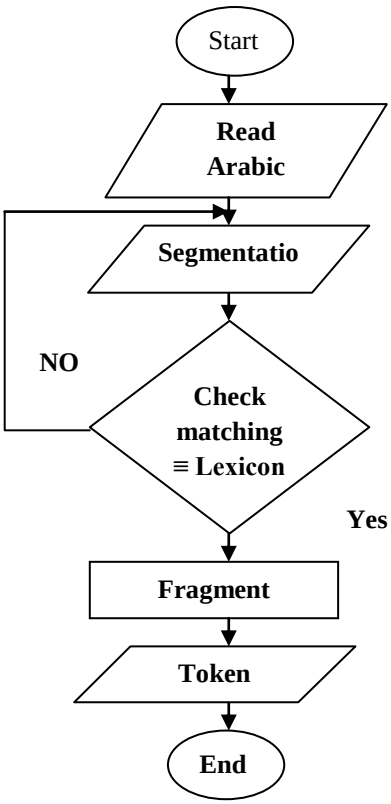


Figure (6) Arabic Tokens Fragments

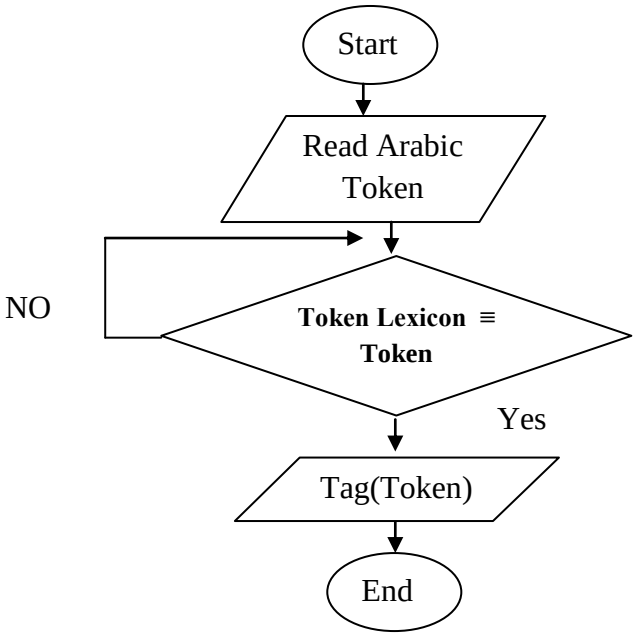


Figure (7) Tagging with lexicon

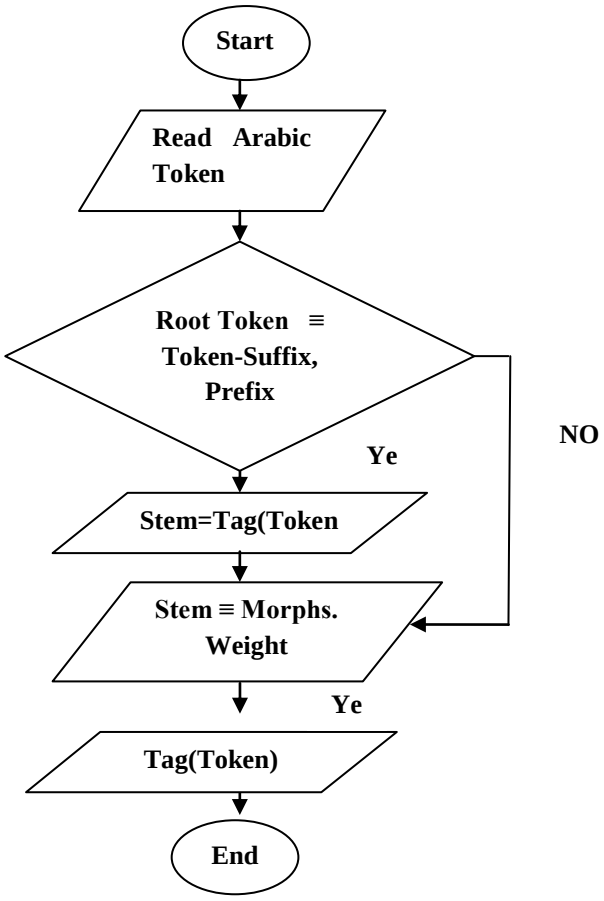


Figure (8) Morphological analysis

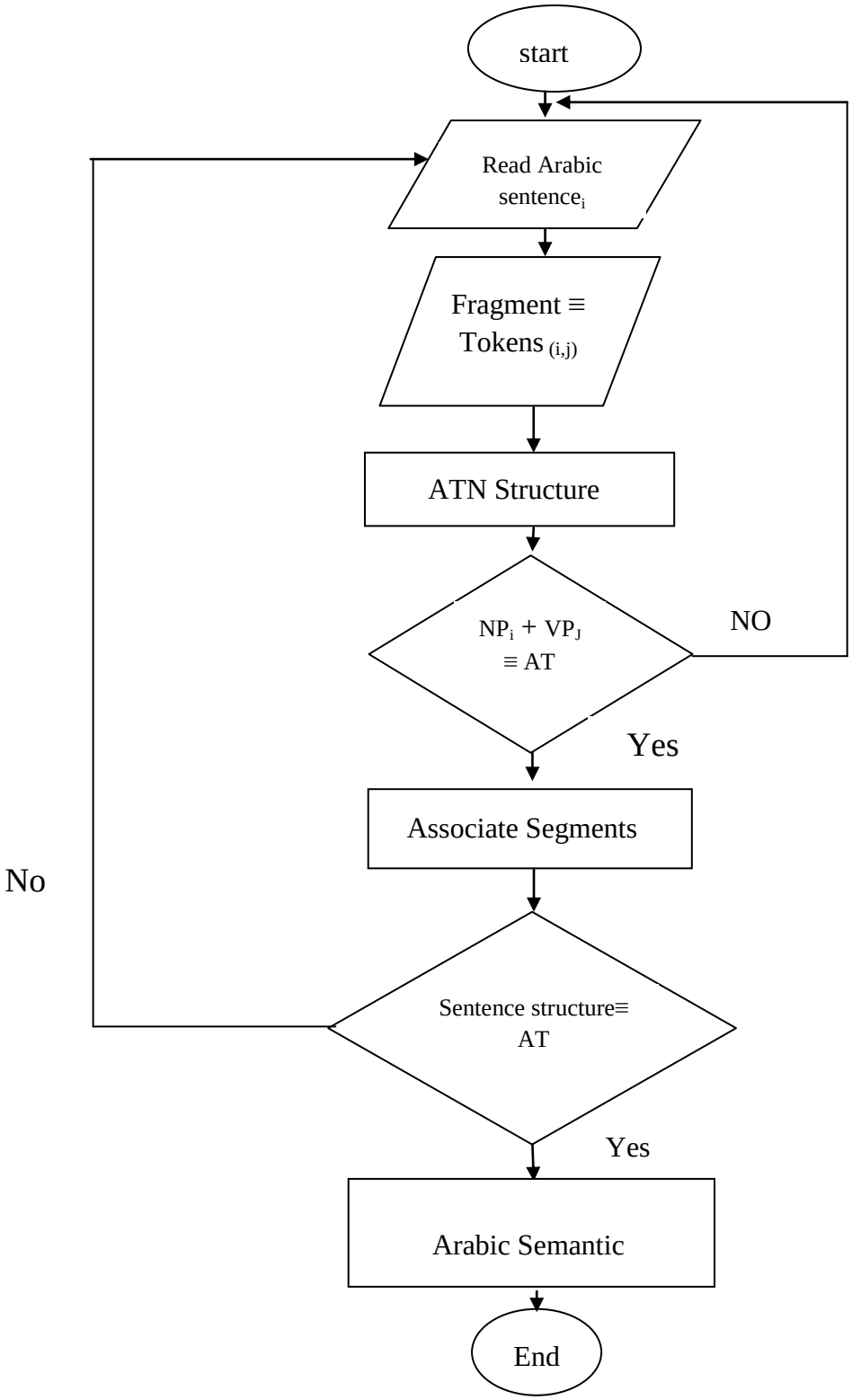


Figure (9) Syntax and Semantic Arabic sentence analysis

Then proposed system train the classifier with words features via adapting Rule-based technique: assigning tags to each token in the sentence after segmentation. Training Model focus on Token features selection after sentence segmentation and Arabic token frequency in these sentences (number of times that a token appears in the processed sentence) as shown in equation (2).

$$\text{TF-IDF}_{ij} = tf * \log\left[\frac{N}{df_i}\right] \ldots\ldots. (2)$$

Where TF-IDF_{ij} is the ith weight of the token in the sentence; TF_{ij} is the number of tokens that the ith token appears in the jth sentence, N is the total number of processed sentences, df_i is the number of sentences in which the ith token appears.

Processing sentences achieved via Turbo Prolog Version 2 on personal computer Pentium 4 with 1.7GHZ processor and 256 RAM. The materials corpus was Arabic sentence for not specific domain.

9. Evaluation

Table [1] shows Experimental results of tokenization precision and Recall of number of sentences. Tested Applied on Unknown word Types.

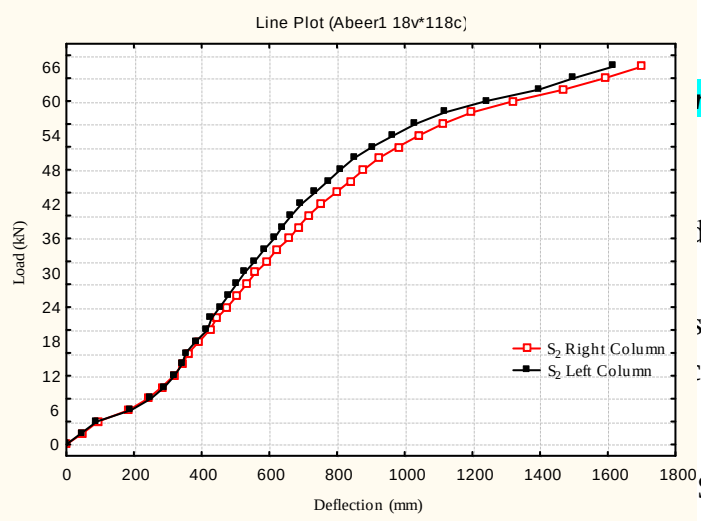
Table [1] Tokenization Precision and Recall

No. of Sentences	Precision	Recall	Fscore
100	82.24	68.73	74.88
200	84.29	77.98	81.01
300	85.88	83.65	84.75
400	87.66	85.14	86.38
500	89.21	87.75	88.47
600	94.75	90.22	92.42
700	95.01	92.16	93.56
800	96.11	94.55	95.32
900	96.87	95.87	96.36
1000	97.51	96.31	96.90

Table [2] presents comparison results of Part of speech tagging accuracies of the proposed system on Arabic words and all test data against memory based approach, SVM-Tokenization, Rule and Rule with Dictionary on Arabic tokenization and Tree bank.

Table [2] Pos Tagging comparison results

System	precision	recall	F-score
Memory based	74.75	81.07	77.78
Rule	86.28	91.09	88.62
Rule+Dict	93.72	93.71	93.71
SVM	93.72	93.71	93.71
Proposed System	97.51	96.31	96.90



l," An Efficient Recursive Transition Network for world congress on engineering,July 1-3 2009.

stical Part-of-Speech Tagger for Traditional Arabic e 5 (11): 794-800, 2009 ISSN 1549-3636 ©2009

Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing. 2nd Edn., Prentice Hall, ISBN: 10: 0131873210, pp: 1024.

4. Alansary, S., M. Nagi and N. Adly, 2008. Towards analyzing the International Corpus of Arabic (ICA). Proceeding of the 8th International Conference on Language Engineering, Egypt.
5. Bosch, Antal, and Erwin, Marsi*, and Soudi, Abdelhadi,"Memory-based morphological analysis and part-of-speech tagging of Arabic", 2004.
6. Guiassa, Yamina," Hybrid Method for Tagging Arabic Text", journal of computer science:2(3):245-248,2006,ISSN 1549- 3636.
7. Jakub, Z. And, Daelemans, W ,2000. "Recent Advances in Memory-Based Part-of-Speech Tagging", Induction of Linguistic Knowledge TSL, 2000.
8. Maamouri, M. and Cieri, C., 2002. Resources for Arabic natural language processing at the LDC. Proceeding of the International Symposium on the Processing of Arabic, Tunisia, 2002, pp: 125-146.
9. Dror, J., D. Shaharabani, R. Talmon and S. Wintner, 2004. Morphological analysis of the Quran. Literary Linguist. Computer. 19: 431-452. DOI: 10.1093/llc/19.4.431
10. Talmon, R. and S. Wintner, 2003. "Morphological tagging of the Quran", Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, Apr. 2003, Budapest, Hungary, pp:1-8. http://cs.haifa.ac.il/~shuly/publications/talmonwintner-eacl03.pdf
11. Al Daoud Essam, and Basata Abdullah," A framework to Automate the Parsing Arabic Language Sentences",Faculity of Science and Information Technology, Zarqa Private University, Jordan, 2005.
12. El hadj, y.o. mohamed, al-sughayeir, and al-ansari, a.m.," Arabic part-of-speech tagging using the sentence structure", centre of research at the college of computer & information sciences , college of Arabic language , imam university , 2007.